

ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

Adam A. Margolin^{1,2}, Ilya Nemenman², Katia Basso³, Ulf Klein³, Chris Wiggins^{2,4}, Gustavo Stolovitzky⁵, Riccardo Dalla Favera³, Andrea Califano^{1,2,3,}*

¹Department of Biomedical Informatics, Columbia University, 622 West 168th Street Vanderbilt Clinic, 5th Floor New York, New York 10032

²Joint Centers for Systems Biology, Columbia University, 1150 St Nicholas Ave, Rm 121, New York, NY 10032

³Institute of Cancer Genetics, Columbia University, Russ Berrie Pavillion, 1150 St Nicholas Ave, New York, NY 10032

⁴Department of Applied Physics and Applied Mathematics; 500 W 120th Street, New York NY 10027

⁵IBM Computational Biology Center, Functional Genomics and Systems Biology Group, IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, N.Y. 10598

Running Head: ARACNE

* To whom correspondence should be addressed

ABSTRACT

Motivation: Cellular phenotypes are determined by the dynamical activity of networks of co-regulated genes. Elucidating such networks is crucial for understanding normal cell physiology and complex pathologic phenotypes, but existing methods for the “reverse engineering” of genetic networks from microarray expression data have been successful only for lower eukaryotes with simple genomes. Here we present *ARACNE*, a novel approach designed specifically to scale up to the complexity of transcriptional regulatory networks in mammalian cells, yet general enough to address a much wider range of network deconvolution problems. This method uses the data processing inequality to eliminate the vast majority of indirect interactions inferred by pairwise information-theoretic analysis.

Results: We prove that the reconstructed topology is exact for large enough data sets if the underlying interactions form a tree, and we show that the algorithm works well in practice, even in the presence of loops and complex network topologies. We assess ARACNE’s ability to reconstruct transcriptional regulatory networks using synthetic data and large-scale microarray profile data from human B cells. ARACNE significantly outperforms both Relevance Networks and Bayesian Networks on the synthetic datasets, while achieving extremely low error rates. Application to the deconvolution of genetic networks in human B cells demonstrates ARACNE’s ability to infer a significant number of putative transcriptional targets of the c-MYC proto-oncogene.

Availability: ARACNE is implemented in the BioWorks platform, which is freely available at: <http://amdec-bioinfo.cu-genome.org/html/BioWorks.htm>.

Contact: califano@c2b2.columbia.edu

Supplementary Information: Supplementary Information and all data used in this paper are available at: <http://www.c2b2.columbia.edu/research/supplemental/aracne-bioinformatics.html>.

1 INTRODUCTION

1.1 Biological Background and Significance

Cell phenotypes are determined by the differential expression of thousands of genes and their products – an activity choreographed by a complex network of interactions that control common functions, such as the formation of transcriptional complexes or the availability of signaling pathways. Thus it is becoming increasingly clear that cellular phenotypes, and the complex range of mechanisms determining their selection, cannot be fully understood unless the function of the individual genes is elucidated in the context of the networks in which they operate. Identifying this organization is especially crucial to dissecting physiology of pathological cells, such as cancerous ones, where the alterations of multiple oncogenes and tumor suppressor genes result in profound and complementary dysregulations of normal cellular pathways.

Genome-wide clustering of gene expression profiles (Eisen, Spellman et al. 1998; Tamayo, Slonim et al. 1999) provides an important first step towards the identification of cellular networks. However, the organization of genes into co-regulated clusters is too coarse a representation to provide clues towards the identification of individual interactions. This is because as biochemical signals travel through cellular networks the expression of many genes that interact only indirectly may become strongly correlated, as has been shown for Cyclin D1 and E2F targets (Lamb, Ramaswamy et al. 2003). More generally, as has long been recognized in statistical physics, a long range order (that is, a high correlation among indirectly interacting random variables) can easily result from only short range, pairwise interactions (Ma 1985). Thus one cannot use correlations, or *any other* local dependency measure, as a tool for the reconstruction of interaction networks without additional assumptions.

Within the last few years a number of more sophisticated approaches for the reverse engineering of cellular networks (also called deconvolution) from gene expression data have emerged. The goal of such methods, briefly stated, is to produce a high-fidelity representation of the cellular network topology as a graph, where genes are represented as nodes and direct regulatory interactions as edges connecting the nodes. While scores of different methods have been proposed [for a review see (Rice and Stolovitzky 2004)], a broad taxonomical organization suggests four major categories. The first includes optimization methods based on the maximization of a high-dimensional objective function associated with different network topologies, such as Bayesian networks (Hartemink, Gifford et al. 2001; Friedman 2004) or Chain Functions (Gat-Viks and Shamir 2003). A common objective function is the log-probability of the network topology given the observed data. The second category includes a variety of regression techniques to fit the observed data to an empirical a-priori model of the underlying biochemical interactions (de la Fuente, Brazhnik et al. 2002; Gardner, di Bernardo et al. 2003; Tegner, Yeung et al. 2003). A third group includes integrative bioinformatics approaches which combine data from a number of independent clues, such as known protein-protein or protein-DNA interactions (from databases or literature), expression data, or DNA binding motifs (Ideker, Thorsson et al. 2001; Steffen, Petti et al. 2002; Middendorff, Kundaje et al. 2004). The fourth category includes statistical/information theoretical methods (Butte and Kohane 2000; Rice, Tu et al. 2004), which define two-

way or higher-order probabilistic measures of gene correlation to distinguish potential interactions from background noise.

Overall, all available approaches suffer from one or more limitations including overfitting (common to regression or optimization methods); exponential complexity and reliance on non-realistic network models (typical of most optimization methods, including Bayesian Networks); or a critical dependency on data that are only available for simple organisms (as is the case for most integrative methods). These limitations have relegated the successful application of such methods to organisms with relatively simple genomes, such as the yeast *Saccharomyces cerevisiae*. As of today, there is no reported example of the application of reverse-engineering algorithms to the genome-wide deconvolution of complex mammalian networks.

This paper introduces a novel, information-theoretic algorithm called ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) for the reverse-engineering of transcriptional regulatory networks from microarray expression data. Using a realistically implemented simulated dataset, we show that ARACNE compares very favorably with existing methods, especially with respect to false positive rates (which are critical in guiding further biological validation experiments). We validate this finding biologically by demonstrating ARACNE's ability to identify putative transcriptional targets of the c-MYC proto-oncogene by analyzing a large-scale set of microarray expression profiles from human B cells.

Although we present this method in the context of transcriptional regulation, we note that this is just one (albeit very important) example of a problem where interaction graphs must be inferred from experimental samples. Other such problems include protein interaction networks (Giot, Bader et al. 2003; Iossifov, Krauthammer et al. 2004), social networks (Barabasi, Jeong et al. 2002), graphical models for representing uncertainty in learning (Pearl 1988), models in statistical physics (Kabashima and Saad 2001; Mezard and Parizi 2001), and many others. The ARACNE algorithm is general enough to be applicable to this entire class of problems.

1.2 Theoretical Background

Several factors have impeded existing methods from reliably reconstructing the complex genetic networks of mammalian cells using currently available data and computation. First, for higher eukaryotes, biological manipulations are difficult and costly. Thus, at best, different cell populations harvested from different individuals capture random steady-states of the underlying biochemical dynamics. The absence of time series data precludes usage of methods like (Wiggins and Nemenman 2003) and others reviewed therein to infer temporal associations and thus plausible causal relations among genes. With little temporal data or specifically targeted biological experiments, only steady-state statistical dependences can be studied, which are not obviously linked to the underlying physical dependency model.

Compounding this constraint is a second problem: there is no universally accepted definition of statistical dependencies in the multivariate setting [See (Joe 1997; Nemenman 2004) for a review of alternatives]. In this work we adopt the definition introduced in (Nemenman and Tishby *Submitted*), which builds on ideas which have been discussed in the literature on Markov networks for several years (Pearl 1988; Kabashima

and Saad 2001). Briefly, by analogy with statistical physics, we write the joint probability distribution (JPD) of the stationary expressions of all genes, $P(\{g_i\})$, $i = 1, \dots, N$, as:

$$P(\{g_i\}) = \frac{1}{Z} \exp \left[-\sum_i \phi_i(g_i) - \sum_{i,j} \phi_{ij}(g_i, g_j) - \sum_{i,j,k} \phi_{ijk}(g_i, g_j, g_k) - \dots \right] \equiv e^{-H(\{g_i\})} \quad (1)$$

where N is the number of genes, Z is the *partition function*, $\phi_i(g_i)$ are *potentials*, and $H(\{g_i\})$ is the *Hamiltonian* that defines the system's dynamics. Within such a model, we can assert that a set of variables interacts if and only if the single potential that depends exclusively on these variables is nonzero. ARACNE aims precisely at identifying which of these potentials are nonzero and eliminating those that are zero even though their corresponding marginal JPDs are significantly non-uniform.

Note that the expansion in Eq. (1) does not define the potentials uniquely, and additional constraints are needed to avoid the ambiguity. The most natural choice (Nemenman 2004) is to determine ϕ_{\dots} by means of the maximum entropy approximations (Janes 1957) to $P(g_1, \dots, g_N)$ consistent with various marginals, so that constraining an n -way marginal defines its corresponding potential. For example, $\phi_i(g_i) = -\log P(g_i)$, but higher order potentials generally cannot be written in a closed form. We refer the reader to (Nemenman 2004; Nemenman and Tishby *Submitted*) for details.

Additionally, we note that, by minimizing the Hamiltonian, the formulation in Eq. (1) allows us to design principled approaches to simulate the steady state behavior of the cellular network from its potentials. As discussed in the next section, depending on the level of approximation, it may be quite feasible to reconstruct the potentials from the marginal JPDs (which can be estimated from the experimental samples) under the maximum entropy approximation. This approach will be explored in future publications.

1.3 Approximations: The interaction structure

Given the relatively small number of microarray samples realistically obtainable, it is infeasible to infer the exponential number of potential n -way interactions suggested by the expansion in Eq. (1). Rather, a set of simplifying assumptions must be made about the variable dependency structure. Eq. (1) provides a principled and controlled way to introduce such approximations. For instance, the simplest model is one where genes are assumed independent, i.e., $H(\{g_i\}) = \sum_i \phi_i(g_i)$, such that first-order potentials can be evaluated from the marginal probabilities, $P(g_i)$, which are in turn estimated from experimental observations. As more data become available we should be able to reliably estimate higher order marginals and incorporate the corresponding potentials progressively, such that for $M \rightarrow \infty$ the complete form of the JPD is restored, where M is the number of experimental samples. In fact, $M > 100$ is generally sufficient to estimate 2-way marginals in genomics problems, while $P(g_i, g_j, g_k)$ requires about an order of magnitude more samples. Thus ARACNE truncates Eq. (1) at the pairwise interactions only, $H(\{g_i\}) = \sum_i \phi_i(g_i) + \sum_{i,j} \phi_{ij}(g_i, g_j)$.

Within this approximation, ARACNE declares two genes to be non-interacting if they are statistically independent (i.e. $P(g_i, g_j) \approx P(g_i)P(g_j)$), and does not model more complex interactions. However, the algorithm further identifies gene pairs that, even though non-independent, have a corresponding zero potential, thus removing the vast majority of false positive interactions.

This formulation is reminiscent of spin glasses on random networks (Mezard and Parizi 2001; Yedidia 2001)}, particularly if the g_i are binary. In this case, the genes are the Ising spins, and truncations to the first, second, or the third order potentials are steps towards the mean field, Bethe, and Kikuchi variational approximations (Bethe 1935; Kikuchi 1951; Oppen and Winther 2001; Yedidia 2001). An important distinction is that in physics one searches for $\tilde{P}(\{g_i\})$, a variational approximation to the true JPD, $P(\{g_i\})$, that minimizes $D_{KL}(\tilde{P}\|P) \equiv \langle \log \tilde{P}/P \rangle_{\tilde{P}}$ within a given class of \tilde{P} , while the definition of (Nemenman 2004) is equivalent to minimizing $D_{KL}(P\|\tilde{P})$. This is because statistical physics aims at calculating various spin statistics given an interaction network. In particular, low order marginals P_{\dots} are unknown and cannot be used in averaging. On the other hand, we are here solving the inverse problem – reconstructing the network given the known marginal distributions.

1.4 Approximations: The network topology

Even considering only pairwise interactions, the problem of reverse engineering the network is still nontrivial. For example, consider the case in which $P(\{g_i\})$ is a multivariate Gaussian. The full joint distribution is then specified by the inverse of the covariance matrix c_{ij}^{-1} . In genomic applications, c_{ij}^{-1} is expected to be sparse; that is, relatively few genes interact directly. However, if $c_{ij} \neq 0$ and $c_{jk} \neq 0$, then generically $c_{ik} \neq 0$. Thus experimentally measured c_{ij} will not reflect the true topology of the network interactions. Furthermore, inversion of large sparse matrices is quite sensitive to small errors; thus the inverse of the measured c_{ij} will not be sparse and simple thresholding will not lead to an accurate reconstruction of the network topology. The problem is even more complex for non-Gaussian distributions (which are far more realistic in actual cellular networks, where interactions are generally non-linear), where mutual information (see Section 2.1) replaces the c_{ij} as the measure of statistical similarity. Since the number of potential pairwise interactions is quadratic in the number of genes, this presents a formidable challenge to all network reconstruction methods, which generically suffer from false positives, as is the case, for instance, for the Relevance Networks approach.

To date, no method has been proposed to solve this issue exactly and to reconstruct an arbitrary two-way interaction network reliably from a finite number of samples in a *computationally feasible time*. However, if the regulatory network can be represented as a tree, an algorithm that can solve the problem exactly is computationally tractable. In fact, as will be shown in the next section, the ARACNE algorithm can “reverse engineer” such tree networks exactly in polynomial time in the limit $M \rightarrow \infty$, when statistical fluctuations in estimating pairwise marginals are small. The method works for marginals

of any form. Furthermore, the algorithm is robust to violations of the tree assumption as long as a local tree-like structure is dominant. This is because nodes in a network generally decorrelate rather quickly, and interactions over more than a few separating edges are generically weak, reducing the impact of large loops. This is a good approximation for biological networks, which are believed to be sparse; consequently, the average size of a random loop is greater than a few genes, unless some yet unknown evolutionary pressure prefers tighter control loops [a notable exception is the feed forward loop, found to be over-represented in biological circuits (Mangan and Alon 2003)]. Thus, as will be shown both from synthetic and from experimental data, ARACNE is very successful in reconstructing complex networks with a large number of loops, even from relatively small sample sizes. We note that other methods can, in principle, reconstruct tree topologies. ARACNE's advantage derives, in particular, from its *provably* exact reconstruction of tree topologies, low computational complexity, small data size requirements, and *robustness* to violations of the tree assumption.

2 ALGORITHM

ARACNE relies on a two-step process. First, candidate interactions are identified by estimating pairwise gene-gene mutual information $I(g_i, g_j)$, an information-theoretic measure of relatedness, and by filtering them using an appropriate threshold, I_0 , computed for a specific p-value, p_0 , in the null-hypothesis of two independent genes. This step is basically equivalent to the Relevance Networks method (Butte and Kohane 2000), and, as such, suffers from critical limitations. In particular, as discussed, while physical interaction will likely result in co-regulation, the reverse is not true. That is, genes separated by one or more intermediaries (indirect relationships) may be highly co-regulated without implying physical interaction, giving rise to false positives.

Thus in its *second step*, ARACNE removes the vast majority of indirect candidate interactions using a well-known property of mutual information, the data processing inequality (DPI), that has not been previously applied to the reverse engineering of genetic networks. Introduction of the DPI produces a dramatic difference in the performance of the algorithm and, in particular, results in a remarkable reduction of false positive interactions with minimal impact false negatives.

We now discuss some aspects of the algorithm, present proofs related to its performance, and then proceed to the results of its applications to natural and synthetic networks.

2.1 Mutual Information

Mutual information (MI) for a pair of random variables, x and y , is defined as

$$I(x, y) = S(x) + S(y) - S(x, y), \quad (2)$$

where $S(t)$ is the entropy of an arbitrary variable t . For a discrete variable the *entropy* is the average of the log-probability of the states:

$$S(t) = -\langle \log p(t_i) \rangle = -\sum_i p(t_i) \log p(t_i) \quad (3)$$

where $p(t_i) = \text{Prob}(t = t_i)$ is the probability of each discrete state (value) of the variable, and logarithms are binary. For continuous variables the entropy is infinite, but the mutual

information remains well defined and can be computed by replacing $S(x)$ with the *differential entropy*, which differs from Eq. (3) in that the log-probability density (rather than the log-mass) is averaged.

Like more familiar correlation measures (e.g. Pearson correlation), MI measures the degree of statistical association between two variables. For instance, for normally distributed covariates it is related to the Pearson correlation. However, while other correlations may be zero even for manifestly dependent variables (such as between a variable symmetrically distributed around zero and its square), MI is guaranteed to be nonzero *iff* any kind of statistical dependence exists. In fact, it is the only measure of statistical association between two variables that satisfies some simple axioms (Nemenman and Tishby *Submitted*). Additionally, MI possesses some critical properties, like the DPI (discussed later), that make it especially desirable for use in network reconstruction applications.

Estimation of mutual information from finite data sets has been extensively studied in the literature (Beirlant, Dudewicz et al. 1997; Strong, Koberle et al. 1998; Nemenman, Shafee et al. 2002; Kraskov, Stoegebauer et al. 2004; Nemenman, Bialek et al. 2004). In general, assuming that the JPDs associated with individual gene pairs are relatively smooth (a realistic hypothesis since system and measurement noise significantly smooth the data), MI can be estimated reliably from a relatively small number of samples (~ 100) using a number of methods. We chose a computationally efficient Gaussian Kernel estimator (Beirlant, Dudewicz et al. 1997). Given two measurement vectors $\{x_i\}$ and $\{y_i\}$, the estimator computes:

$$I(\{x_i\}, \{y_i\}) = \frac{1}{M} \sum_i \log \frac{f(x_i, y_i)}{f(x_i)f(y_i)} \quad (4)$$

where $f(x, y)$ and $f(x)$ are Gaussian kernel density estimators defined as:

$$f(x, y) = \frac{1}{2\pi h^2 M} \sum_i \exp \left\{ -\frac{(x - x_i)^2 + (y - y_i)^2}{2h^2} \right\}, \quad f(x) = \frac{1}{\sqrt{2\pi} h M} \sum_i \exp \left\{ -\frac{(x - x_i)^2}{2h^2} \right\},$$

where $h = h(M)$ is the kernel width. This estimator is asymptotically unbiased for $M \rightarrow \infty$, as long as $h(M) \rightarrow 0$ and $[h(M)]^2 M \rightarrow \infty$, e.g. $h(M) \propto M^{-1/t}$, $t > 2$.

Gaussian Kernel Width: While the estimator is asymptotically unbiased for any $h(M)$ that satisfy the above conditions, the choice of $h(M)$ that removes the bias for a finite M is critically dependent on M and on the JPD being analyzed, specifically on its smoothness. In general, as h grows, the kernel density estimator becomes almost uniform and MI is underestimated. In contrast, small h produces peaked distributions with overestimated MI. This is illustrated in Figure 1, which shows that the average absolute error on the MI estimate, $\langle |I - \bar{I}| \rangle$, is highly sensitive to the choice of the kernel width (here \bar{I} is the true value of the MI, and I is its estimate). While good methods for determining the kernel width exist [see e.g. (Nemenman and Bialek 2002)], it is computationally costly to apply such analysis to each gene pair. Fortunately, ARACNE's performance does not depend directly on the accuracy of the MI estimate, but rather on

the accuracy of the estimation of MI differences. For instance, determining if MI is statistically significant requires testing whether $\bar{I}_{ij} \geq I_0$, where $\bar{I}_{ij} = \bar{I}(g_i, g_j)$ and I_0 is the threshold for statistical significance computed from Monte Carlo. Similarly, the DPI (Section 2.3) only requires ranking the mutual informations, for which the estimate of their differences is also sufficient.

Producing reliable estimates of the MI differences, $\bar{I}_{ij} - \bar{I}_{kl}$, is a much easier task because the bias tends to cancel out, especially for $\bar{I}_{ij} \approx \bar{I}_{kl}$, which corresponds to the most error-prone cases in ranking relative MIs. In particular, selecting a single “ensemble best” value of h (see below) for the data in question, we expect that the estimated MI may be biased. That is, $I = \bar{I} + b + \varepsilon$, where b is the bias, and ε is the zero-mean random fluctuation. From the work on estimation of MI for discrete variables (Strong, Koberle et al. 1998), we expect the bias to be a function of marginal and joint entropies. Thus for well-sampled marginals and an undersampled joint, we have $b \approx b(\bar{I}, h)$, and the biases cancel out for similar values of MIs. This is especially true since, due to the same underlying noise sources, the JPDs for various gene pairs have similar smoothness properties.

Figure 1 presents a numerical analysis in support of this argument. The solid brown curve shows the average number of errors in ranking pairs of MIs. Indeed, the minimum of the function is very broad, as opposed to that for the MI estimate, and rank ordering of MI is rather insensitive to the kernel width. That is, even when \bar{I} s are uncertain, the order of the estimates remains stable. Thus instead of selecting a different h for each gene pair, we settle for a single value that minimizes $\langle |I - \bar{I}| \rangle$ for the Gaussian distributions shown in Figure 1 for an equivalent number of samples as in the data in question. In Section 3.3.3 we demonstrate that this choice of the kernel width largely corresponds to the optimal value for reconstruction of the synthetic network. While, in view of this discussion, uniform h results only in a minimal performance loss, in future publications we plan to explore determining h for individual gene pairs to further increase the accuracy of our estimates.

2.2 Statistical Threshold for Mutual Information

Since the mutual information is positive semi-definite, its evaluation from random samples gives a positive value even for variables that are, in fact, independent. Therefore, we eliminate all edges for which the null hypothesis of independent genes cannot be ruled out with a given certainty. To this extent, we randomly shuffle the expression of genes across the various microarray profiles, similar to (Butte and Kohane 2000). We then evaluate the MI for such manifestly independent genes and empirically estimate the fraction, p , of the estimates above some threshold value I_0 . Inverting the relationship tells us which MI estimates are below the threshold and which corresponding edges are to be eliminated if we require a p-value not greater than p . This is done for different sample sizes M and for 10^5 gene pairs so that reliable estimates of $I_0(p)$ are produced up to $p = 10^{-4}$. Extrapolation to smaller p-values is done using:

$$p(I \geq I_0 | \bar{I} = 0) \propto e^{-\alpha MI_0}, \quad (5)$$

where the parameter α is fitted from the data. This formula is based on the intuition of the large deviation theory (Cover and Thomas 1991), which for discrete data and unbiased estimators suggests $p(I \geq I_0 | \bar{I} = 0) \propto e^{-MI_0}$. As mutual information in the continuous case can be estimated by finely discretizing the variables, a similar result should hold. We then use the parameter α to account for possible biases of the estimator at fixed h , producing excellent agreement with numerical experiments (see supplemental).

2.3 Data Processing Inequality

The DPI (Cover and Thomas 1991) states that if genes g_1 and g_3 interact only through a third gene, g_2 , [i.e., if the interaction network is $g_1 \leftrightarrow \dots \leftrightarrow g_2 \leftrightarrow \dots \leftrightarrow g_3$ and no alternative path exists between g_1 and g_3], then the following inequality holds

$$I(g_1, g_3) \leq \min[I(g_1, g_2); I(g_2, g_3)]. \quad (6)$$

Thus under some circumstances the least of the three MIs can come from indirect interactions only. Correspondingly, ARACNE starts with a network graph where each $I(g_i, g_j) > I_0$ is represented by an edge (ij) . The algorithm then examines each gene triplet for which all three MIs are greater than I_0 and removes the edge with the smallest value (see Figure 2). We emphasize that each triplet is analyzed irrespectively of whether one of its edges has been marked for removal by a prior DPI application to a different triplet. Thus the network reconstructed by the algorithm is independent of the order in which the triplets are examined.

Theorem 1. If MIs can be estimated with no errors, then ARACNE reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.

Proof of Theorem 1. First, notice that for every pair of nodes g_i and g_k not connected by a true direct interaction there is at least one other node g_j that separates them on the network tree. Applying the DPI to the (ijk) triplet leads to removal of the (ik) edge. Thus only true edges survive. Similarly, every removed edge is not present in the true network. Consider some (ijk) triplet. One of its genes, say g_j , may separate the other two. In this case the removed edge (ik) is clearly not in the true tree. Alternatively, there may be no separating gene, and one may be able to move between any gene pair in the triplet without going through the third one. In this case none of the three edges is in the true graph, and any edge the DPI removes is fictitious. Thus all removed edges are indirect, while all remaining edges are factual. The network is reconstructed exactly.

The algorithm is not guaranteed to reconstruct correct networks if loops are present (in fact, *every* loop with only three genes will be opened along the weakest edge). However, if loops are large, then locally the network looks like a tree. Thus, as in the corresponding discussion in statistical physics (Yedidia 2001), algorithms designed for trees still work well. This is because the system is not deterministic, and influence of a gene on another one falls off quickly with the separation between them. Additional long paths between two genes contribute only negligibly to statistical associations coming from direct interactions and are unlikely to impair ARACNE's performance, as will become evident in the results section. In general, we expect that the performance will decrease in the

presence of large numbers of tight loops, especially 3-gene loops, which will be opened by the DPI on their weakest interaction.

Additionally, we note that to minimize the impact of potential MI estimation errors, a tolerance, τ , may be introduced such that the DPI inequalities become of the form $I_{ij} \leq I_{ik}(1 - \tau)$, and close values of MI are not pruned. Using such non-zero tolerance leads to persistence of some 3-gene loops.

2.4 Algorithmic Complexity

Because for a network of N genes there are at most N choose 3 gene triplets, ARACNE's complexity is $O(N^3 + N^2M^2)$, where M is the number of samples and N is the number of genes. The first term relates to the DPI analysis and the second to the mutual information estimate. This compares favorably with optimization methods that must explore an exponential search space. In practice, the DPI is applied to a small subset of triplets for which all three edges survived the mutual information thresholding. Therefore, for large M , the computationally intensive part is generally associated with the second term (computing mutual information), which scales as $O(N^2M^2)$. As a result, ARACNE can efficiently analyze networks with tens of thousands of genes.

3 RESULTS

We study ARACNE's performance on reconstructing networks using three different datasets: a small and well studied galactose metabolism network in *S. cerevisiae*, synthetic networks proposed by (Mendes, Sha et al. 2003), and a large mammalian genetic network inferred from gene expressions of human B lymphocytes. ARACNE's performance is compared against Relevance Networks (RNs) and Bayesian Networks (BNs). RNs are important to characterize the improvement associated with the introduction of the DPI, while BNs have emerged as some of the most widely used reverse engineering methods and provide an ideal comparative benchmark.

3.1 Comparative Algorithms

A *Bayesian Network* is a representation of a JPD as a directed acyclic graph (DAG) whose vertices correspond to random variables $\{X_1, \dots, X_n\}$, and whose edges correspond to parent-child dependencies among variables, see (Pearl 1988) for an introduction and (Heckerman 1999) for a more recent tutorial. We implemented the BN algorithm in this work in accordance with (Hartemink, Gifford et al. 2001; Yu, Smith et al. 2002). In particular, we score graphs using the Bayesian scoring metric (Cooper 1992), for which we adopt a uniform prior over graphs and employ a Dirichlet prior over parameters to aid in the inference of undersampled conditional distributions of children given their parents. Such an approach inherently penalizes more complex graphs. Learning the most likely network requires exploring the entire graph space for the highest scoring model, which is an NP-complete problem (Chickering 1996). Thus heuristic procedures such as greedy hill climbing or simulated annealing are used to search for locally optimal graph structures. The comparative tests presented here use the greedy hill climbing algorithm with random restarts (various structure search methods were tested and observed to produce similar results). BN results were produced using the software from Nir

Friedman’s Computational Biology group at the Hebrew University (Friedman and Elidan 2004), which is among the best implementations of the method.

The other algorithm used for comparison, *Relevance Networks* (Butte and Kohane 2000), computes mutual information for all gene pairs in a microarray dataset and infers that two genes are biologically related if they have the MI above a certain threshold. This approach is basically equivalent to the first step in the ARACNE algorithm, without introduction of the DPI. Therefore, we construct Relevance Networks by running ARACNE with a tolerance on the DPI of 100%. Note that ARACNE uses a different, more accurate mutual information estimator than the one proposed in the original work.

3.2 Yeast Galactose Pathway

We start with a very simple galactose metabolism network in *S. cerevisiae*, which was extensively studied by (Hartemink, Gifford et al. 2001) to test a Bayesian Network approach to genetic network reconstruction. This simple network is useful to highlight some potential pitfalls of Bayesian Networks that are addressed in the context of ARACNE. This network involves three genes, Gal80, Gal4, and Gal2, for which there is an accepted biochemical interaction model: Gal4 activates transcription of Gal2, while Gal80 inhibits Gal4 post-translationally, by protein-protein interaction. Therefore, in the correct model, the expression of Gal2 is determined simultaneously by Gal80 and Gal4.

Eleven distinct Bayesian Networks exist for a such three-gene system. Treating edges as non-directional for comparative purposes collapses the networks into eight topologies (Figure 3.a), whose scores can be evaluated given the data. The Bayesian Network paradigm is that the network most likely to have produced the data will have the highest score. However, in this case, an incorrect model (#1, Gal2 is only transcriptionally controlled by Gal80) has the highest score rather than the correct one (#4, Gal4 and Gal80 jointly control Gal2). Additionally, another model (#6, Gal4 and Gal2 jointly control Gal80) is equiprobable to the correct one, but is also wrong. This shows that even for a toy problem the large number of possible network configurations is an obstacle to the reconstruction of the correct topology. In particular, small sample size, the choice of priors, the choice of discretization boundaries, and the noise can all contribute to an incorrect reconstruction. For instance, the small sample size produces undersampled child-parent conditional distributions even when expressions are discretized to just three levels. More complex networks make the problem exponentially more difficult.

We note, however, that the mutual information between Gal80 and Gal4 is much smaller than that in the Gal2-Gal80 and Gal2-Gal4 edges. Thus RNs can reconstruct the correct topology provided the mutual information threshold is chosen appropriately, while ARACNE does even better by removing the Gal4-Gal80 edge in its DPI step irrespective of the threshold (Figure 3.b).

3.3 Synthetic Networks

3.3.1 Networks Specification: The second series of comparisons uses synthetic transcriptional networks that consist of 100 genes and 200 interactions[†] organized in an

[†] We will evaluate network recovery based on the number of interactions after eliminating auto-regulation and bidirectional edges.

Erdős-Rényi (random network) (Erdos and Renyi 1959) or a scale-free (Barabasi and Albert 1999) topology (see Figure 4). In the former each vertex of a graph is equally likely to be connected to any other vertex, while in the latter the distribution of the number of connections associated with each vertex follows a power law distribution $p(k) \sim k^{-\gamma}$ with $\gamma > 0$. Many real biological networks, including metabolic (Jeong, Tombor et al. 2000), protein-protein (Jeong, Mason et al. 2001), and transcriptional (Babu, Luscombe et al. 2004) ones have been shown to exhibit a scale-free topology.

These synthetic network topologies and dynamics were proposed by (Mendes, Sha et al. 2003) as a realistic platform for comparison of reverse-engineering algorithms because of (a) their realistic complexity, (b) the presence of many regulatory loops, (c) the presence of a few highly interconnected genes (for the scale-free version), and (d) the biologically motivated non-linear transcriptional dependencies among genes. These networks use a multiplicative Hill dynamics (Hill 1910) to model transcriptional interactions.

$$\frac{dx_i}{dt} = a_i \prod_{j=1}^{N_I} \frac{IK_j^{n_j}}{IK_j^{n_j} + I_j^{n_j}} \prod_{l=1}^{N_A} \frac{A_l^{m_l}}{AK_l^{m_l} + A_l^{m_l}} - b_i x_i, \quad (7)$$

where x_i is the concentration (expression) of the i -th gene. N_I and N_A are the number of upstream inhibitors and activators respectively, and their concentrations are I_j and A_l . All other parameters (a_i , b_i , IK_j , AK_l , n_j , and m_l) are specified in (Mendes, Sha et al. 2003).

The synthetic expression value of each gene x_i in each microarray M_k was obtained by simulating this dynamics until the system relaxes to a steady state $\dot{x}_i \approx 0$. For each simulation, the rates of synthesis and degradation were varied by setting $a_i = \lambda_{k,i} \bar{a}_i$ and $b_i = \gamma_{k,i} \bar{b}_i$, where \bar{a}_i and \bar{b}_i are the original constant values of the parameters, and $\lambda_{k,i}, \gamma_{k,i}$ are random variables uniformly distributed in $[0.0, 2.0]$. These parameters are generated once, independently for each gene in each synthetic microarray, and are then kept constant across the entire time-course. Note that $\lambda_{k,i} \sim 0.0$ corresponds to a gene knock-out, while $\lambda_{k,i} \sim 2.0$ corresponds to a 2-fold increase in the synthesis rate. This parameter randomization is intended to model the sampling of a population of distinct cellular phenotypes at random time points (but in equilibrium), as was done for the B cell experiments described later, where the efficiency of individual biochemical reactions may be different from assay to assay due to differences in temperature, nutrients, genetic mutations, etc. Interestingly, under such perturbations, the relationships between simulated genes across all experiments appears very similar to that of real experimental data obtained from the B cells (see supplemental).

3.3.2 Performance metrics: As reverse engineering can be described as classifying each pair of nodes as having or not having an edge, the performance is typically assessed using the same measures as for classifiers: (a) *true* and *false positives*, N_{TP} and N_{FP} – the number of inferred interactions that are present and not present in W , and (b) *true* and

false negatives, N_{TN} and N_{FN} – the number of potential edges correctly and incorrectly identified as not in W . Since genetic networks are believed to be sparse, the number of potential true negatives far exceeds potential true positives. Thus performance metrics traditionally used in ROC analysis, in particular *specificity*, $N_{TN} / (N_{FP} + N_{TN})$, are a bad match for the problem: for all reasonable reconstructions the specificity is close to 1. Therefore, we choose to focus on other performance measures: *precision* and *recall*. Recall, $N_{TP} / (N_{TP} + N_{FN})$, indicates the fraction of true interactions correctly inferred by the algorithm, while precision, $N_{TP} / (N_{TP} + N_{FP})$, measures the fraction of true interactions among all predicted. Precision also provides a meaningful metric as it provides an estimate of the probability that a predicted edge is real, and therefore corresponds to the expected success rate in the experimental validation of predicted interactions.

In accord, we use Precision vs. Recall curves (PRCs) rather than the more conventional ROC curves. PRCs are generated by adjusting some parameter so that N_{TP} goes from 0 to the maximum number of true positives for a method. For ARACNE and RNs this parameter is the p-value or, equivalently, the mutual information threshold. For RNs, $p_0 = 1$ keeps all interactions, leading to 100% recall but a very low precision. On the contrary, even at $p_0 = 1$, ARACNE's DPI eliminates many interactions, leading to a much higher precision. For ARACNE the best recall is ~68% and ~53% at the minimum precision of ~71% and ~38% for Erdős-Rényi and scale-free topologies, respectively. To reach the 100% recall, the DPI tolerance, τ , can be adjusted, until ARACNE's PRC degenerates into that of RNs. For BNs, the adjustable parameter is the Dirichlet pseudocount, and, again, we observe that the maximum recall never reaches 100%. In fact, for either topology the highest recall value achieved by BNs (using a pseudocount of 1,000) is ~44% with a precision of ~14%. Such precision would likely be too low to justify experimental validation of the results for a real biological network.

3.3.3 Performance Evaluation: As shown in Figure 5, PRCs for ARACNE are consistently better than those for BNs and RNs. That is, for any reasonable precision (i.e. > 40%), ARACNE has a significantly higher recall for either topology than the other methods, and its precision reaches ~100% at significant recall values.

The reason for ARACNE's success can be seen by analyzing the distribution of MIs as a function of the length of the shortest path connecting each gene pair (degree of connectivity). ARACNE depends on MI being enriched for directly interacting genes and decreasing rapidly with this distance. Figure 6 shows this to be the case for our simulated datasets: MI is rapidly reduced as the degree of connectivity increases, until its distribution is indistinguishable from the background (Figure 6.b, inset). This highlights two important points. First, there is no unique choice for the MI threshold that separates directly and indirectly interacting genes. As a result, methods such as RNs that attempt to use a single threshold will either recover many indirect connections or miss a substantial number of directly interacting genes. This is obvious from the PRC for Relevance Networks. Second, mutual information decreases rapidly as signals travel over the network, raising the possibility of eliminating a substantial number of distant indirect associations by imposing a slightly conservative threshold that will eliminate only a few

true interactions, while connections with enriched mutual information due to indirectly interacting genes can be eliminated a-posteriori via the DPI. Moreover, because signals in this network decorrelate rather quickly the statistical properties of a tree-like structure will be locally preserved in the presence of loops that contain more than a few genes.

Table 1 shows the number of true and false positives inferred by each algorithm for each network topology and varying synthetic microarray sizes.

Erdős-Rényi networks: Using a sample size of 1,000 synthetic microarrays, ARACNE is able to recover, on average, 128 out of 194 true connections with only 1.3 false positives. As a comparison, RNs recover an average of 143 true connections with 462.7 false positives. Therefore, the DPI eliminates 461 false positives while reducing the number of true positives by only 15, yielding a DPI sensitivity of 99.71%, calculated as the percent of false positives eliminated, and a DPI precision of 96.8%, calculated as the percent of false positives removed out of the total number of edges removed. Bayesian Networks recover an average of 52.7 true connections with 35.3 false positives.

ARACNE's network reconstruction performance is stable as the number of samples decreases. In particular, the number of true positives recovered by ARACNE decays gracefully while the number of false positives remains very low. For a sample size of 125 synthetic microarrays, ARACNE still recovers 81 true connections and 4.3 false connections, with a DPI sensitivity of 95.1% and a DPI precision of 96.1%. The performance of Bayesian Networks degrades rapidly as the number of samples decreases, because the conditional probability tables become very sparsely populated. For a synthetic microarray size of 125, BNs recover an average of 7.3 true connections with 22.7 false connections.

Scale-free networks: Using a sample size of 1,000 synthetic microarrays, ARACNE recovers an average of 97.7 true connections and 2.3 false connections, while RNs recover 113.3 true connections and 234 false connections, corresponding to a DPI sensitivity of 99% and a DPI precision of 93.67%. Bayesian Networks recover an average of 40 true connections with 18.7 false positives.

When the sample size is reduced to 125 synthetic microarrays, ARACNE recovers 46.3 true connections and 3.7 false connections, with a DPI sensitivity of 92.6% and a DPI precision of 96.5%. BNs' performance again deteriorates rapidly at low sample sizes, inferring only 4.3 true connections with 7 false connections.

In general, for all tested sample sizes and for both network configurations, Bayesian Networks recover far fewer true connections and far more false connections than ARACNE. The same is true for Relevance Networks, unless the precision is reduced to below ~35%.

3.3.4 Parameter Estimation: The analysis described in this section for ARACNE and Relevance Networks was performed by fixing the estimator's Gaussian kernel width to the value yielding the $\min \left[\left| I - \bar{I} \right| \right]$ for numerical simulations using an equivalent number of samples drawn from Gaussian distributions, as described in Section 2.1. As shown in Figure 7, this value does in fact largely optimize or nearly optimize the network recovery for ARACNE, as measured by the minimum total number of errors

($N_{FP} + N_{FN}$). Moreover, the network inference is very stable for a large interval of Gaussian kernel widths, verifying the intuition motivated in Section 2.1 that the MI rank error is far more robust to the choice of kernel widths than the MI estimation error.

As shown in Figure 8, the DPI tolerance can be increased up to $\sim 20\%$ with limited impact on false positives, while larger tolerance values produce a much sharper increase. Hence, a moderate choice for the tolerance can help elucidate additional interactions without introducing an excessive number of false positives. This will be used to our advantage in the biological network reconstruction for human B cells.

In summary, ARACNE appears (a) to achieve very high precision and substantial recall, (b) to be stable with respect to the choice of parameters h (Gaussian Kernel width) and I_0 (statistical threshold), and (c) to achieve substantial recall and high precision even with very few data points (125).

3.4 Human B Cells

Removed pending journal publication.

4 LIMITATIONS AND FUTURE WORK

ARACNE drastically improves network inference due to its efficiency in filtering false-positives (see Table 1). However, two issues arise: potential loss of three-way and higher-order interactions, and the opening of three-gene loops. We address each issue separately and offer suggestions for future investigation. We note that the suggestions are compatible with the original formulation of Eq. (1) and correspond to expanding the potentials up to the third order, rather than stopping at the second order.

4.1 Three-Way Interactions

One extension of the current formulation would address the constraint that the statistical filtering will prune all three-way and higher order interactions between genes that cannot be expressed as pairwise interaction potentials. A biological example is that of two transcription factors, g_A and g_B , that independently activate gene g_C , but form an inactive complex, g_{AB} , that fails to activate transcription. This produces an activation pattern akin to an XOR Boolean table. It is well known that for such an interaction the mutual information between any gene-pair is zero; by truncating Eq. (1) at the pairwise interactions, ARACNE would declare these genes statistically independent. However, we note that such idealized situations are quite implausible. Biochemical reactions that produce higher order interactions usually create corresponding lower order dependencies as well. In fact, in (Nemenman 2004) a continuous variables example with *just* third order interactions could not be found. Thus if one is interested only in whether an interaction between a pair of genes is present and does not care for the type of the interaction [the usual approach of Markov networks (Pearl 1988)], then the truncation of the Hamiltonian is not likely to lead to serious systematic errors. We see this in both the galactose network and in the multiplicative Hill dynamics synthetic networks, both of which are reconstructed well by ARACNE in spite of the presence of higher order interactions. However, we emphasize that our formulation, in principle, can distinguish a much richer set of interactions, including those among pairs, triplets, and larger sets of variables, up to

$\sum N!/[(N-i)!i!]=2^N$ different interactions. This is in contrast to alternative approaches, such as Bayesian or Markov networks, that represent statistical interactions as graphs and can only determine which of the $N(N-1)/2$ edges are present. While practical development of relevant algorithms is left for future publications, we expect that studying analogs of trees for higher order interactions (Kikuchi approximation and beyond in statistical physics) will allow us to design DPI-like inequalities capable of recovering such higher order interactions based on multiinformation (Nemenman and Tishby *Submitted*), the analog of mutual information for more than two variables.

An alternative strategy to recover three-way interactions could employ a conditional mutual information:

$$I(g_x, g_y | g_z) = \int I(g_x, g_y | g_z = z) p(g_z = z) dz . \quad (8)$$

For instance, the conditional MI between g_B and g_C , given g_A , in a XOR network is $I(g_B, g_C | g_A) = \log 2$, which is different from zero and allows the interaction to be recovered. We continue the discussion of the conditional MI approach in the next section.

4.2 *Three-Gene Loops*

A second extension of the current formulation would avoid the present constraint that all three-gene loops will be opened along the weakest interaction (although some may be preserved when a non-zero DPI threshold is used at the expense of some additional false positives). Such an extension could attempt to exploit the fact that any edge $g_B \leftrightarrow g_C$ that was correctly removed from the triplet ($g_B \leftrightarrow \dots \leftrightarrow g_A \leftrightarrow \dots \leftrightarrow g_C$) by the DPI should have $I(g_B, g_C | g_A) = 0$. Unfortunately, assessing that the conditional mutual information is zero requires a very large sample size because the MI between g_B and g_C must be estimated over a vanishingly small expression range of the conditional gene g_A . Thus a better approach may be to search for a specific interval, \tilde{A} , of the expression values of g_A for which the DPI is violated, i.e. $I(g_B, g_C | g_A \in \tilde{A}) > I(g_A, g_{B(C)} | g_A \in \tilde{A})$. If such an interval can be found, then the edge $g_B \leftrightarrow g_C$ may need to be reintroduced.

4.3 *Edge Directionality*

A third extension would allow inference of directed edges. We note, however, that this problem is not effectively addressed even by algorithms that formally infer edge directions. For instance, Bayesian Networks inference on the synthetic networks assigned incorrect directionality to about half of the predicted edges. We intend to further study edge directionality using a two-tier approach in which first adirectional gene interactions are inferred, and then edge directionality is assessed via regression algorithms.

4.4 *Extensions to the Validation Framework*

We plan to extend the synthetic networks analysis described in this paper to investigate alternative models that account for more complex *cis*-regulatory dependencies between genes, as well as non-transcriptional dependencies. Additionally, we will explore the role of Langevin noise and detection noise on network reconstruction performance, and the simulation of specific synthetic gene knock-outs in combination with the parameter randomization approach. We also intend to test the network reconstruction degradation as a function of the percent of “hidden” synthetic genes, i.e., genes that are present in the

network but not available to the reverse-engineering analysis. This will allow us to better investigate the issues associated with our limited availability of monitored molecular species. Finally, in addition to the analysis performed on human B lymphocytes, we will apply ARACNE to the deconvolution of well characterized regulatory networks in *S. cerevisiae*.

5 DISCUSSION

Genetic regulatory networks can be described in terms of information flow carried by gene regulatory molecules. Due to inherent stochasticity in biochemical reactions information is lost as a signal propagates through a network. In this paper, we propose an information-theoretic methodology that exploits these characteristics and uses local statistics to infer the most likely path of information flow. We first introduce a formalism that can be used to represent any interaction network, not limited to pairwise interactions. We then proceed to justify a set of simplification rules that limit the interactions to those that can be reliably inferred from experimental data sets. Based on this representation we propose an algorithm, ARACNE, that can exactly infer tree-like networks, and we show, by validation against other methods on realistic-complexity synthetic networks, that ARACNE works extremely well even in the presence of many tight loop structures. This method extends upon traditional clustering based approaches and reconstructs more intricate dependencies within gene clusters. It also overcomes some critical limitations of optimization methods, such as Bayesian Networks, because it has low computational complexity, does not require discretization of the expression levels, and enables the reconstruction of larger loops. ARACNE can be applied to arbitrarily complex networks of transcriptional interactions without reliance on heuristic search procedures. Thus it is ideally suited for mammalian gene regulatory networks which (a) are characterized by a complex topology, (b) do not benefit from well-defined supplemental data (such as comprehensive protein interaction databases available for yeast), and (c) are more difficult to manipulate experimentally, substantially hindering the acquisition of data to which time-series based methods can be applied.

We tested this method on a mammalian network by analyzing a large panel of microarray expression profiles from human B cells that span a substantial phenotypic variety. This approach differs from traditional methods that rely on systematic perturbations to simple organisms, which are not easily performed in mammalian cells. Using this data, ARACNE is able to construct a highly complex network with 129,000 interactions. Analysis of the network structure surrounding the c-MYC proto-oncogene reveals a significant enrichment in bona-fide c-MYC targets, and application of the DPI is shown to be effective in identifying direct targets of c-MYC by literature analysis and biochemical validation.

We also thoroughly benchmarked ARACNE against other reverse-engineering algorithms (i.e., Bayesian Networks and Relevance Networks) using a realistically implemented synthetic simulated dataset designed to approximate the steady-state dynamic richness of expression profiles obtained by sampling different phenotypes. We examined two alternate topologies for these data: the Erdős-Rényi (or random network), which assumes no prior knowledge of the topological structure, and the scale-free topology, which approximates some more complex features of biological networks. The latter presents

greater challenges for reconstruction due to the presence of many small loops and multiple regulators that obscure direct dependencies between gene pairs. For example, a random network with ~ 2 connections per node produces an average loop size $\sim \sqrt{N} \approx 10$, so the network is locally a tree. On the other hand, small loops are common in the scale-free network. As ARACNE is exact for trees, its performance on Erdős-Rényi networks is somewhat better than on the scale-free ones. Interestingly, Relevance Networks produce lower false-positive rates on the scale-free network. This is because indirect interactions are more likely to pass through a large interaction hub, which, due to their large in-degree, decorrelate signals much faster than poorly connected genes. This is evident by observing that the distribution of MIs for genes that are further than third neighbors in the scale-free networks already roughly approximates the background distribution (Figure 6.b, inset), whereas many genes separated by an equivalent path length in the Erdős-Rényi topology have statistically significant MI (Figure 6.a). Thus a much larger number of indirect connections can be eliminated in the scale-free networks by imposing a statistical significance threshold.

An additional factor compounding the reconstruction of the synthetic scale-free networks might be due to a biologically implausible simplification of the synthetic model: the networks used here contain hubs with very large numbers of inbound connections (large in-degree), while biologically we expect a hub gene to regulate many other genes (large out-degree). High in-degree effectively masks pairwise interactions, decreasing the MI and causing degraded performance of ARACNE. Since larger sample sizes are necessary to accurately estimate MI for high in-degree nodes, the number of true positives inferred by both ARACNE and RNs decays more rapidly with decreasing sample sizes for the scale-free topology than the Erdős-Rényi topology. While these topological differences affect the estimation of pair-wise MI and with that the number of true positives, in spite of these confounding factors of the network topologies, and for all numbers of samples, the DPI is able to eliminate nearly all false interactions that were inferred by Relevance Networks at the expense of very few true interactions, indicating considerable robustness. However, as the p-value increases above reasonable values for a network of this size (i.e. $\sim 10^{-4}$), ARACNE begins retaining a large number of false candidate interactions without any additional increase in true positives, as is evident by the dramatic drop in precision for very high p-values (the right tail of ARACNE's PRC). This is because the DPI may produce random results for very low MI values as small statistical fluctuations may change the rank ordering of mutual information. Therefore, a conservative threshold should be used that eliminates gene pairs with very low MI values, while the DPI can eliminate the vast majority of remaining indirect candidate interactions.

Although the ARACNE is highly accurate in removing false interactions in the simulated dataset, application of the DPI is ill-suited to the inference of certain control structures; in particular, three-gene loops and three-way interactions, and improvements to the algorithm must be investigated to address these conditions. However, in its current instantiation ARACNE has been demonstrated to outperform accepted Bayesian Network and Relevance Network methods, and to be highly effective in eliminating false candidate interactions in a realistically implemented simulated dataset, as well as to identify putative transcription factor targets in human B cells. There are currently no other examples of a genome-wide mammalian network inferred from microarray

expression profiles. As a result, ARACNE shows significant promise in an area that has traditionally challenged reverse engineering algorithms.

6 REFERENCES

- Babu, M. M., N. M. Luscombe, et al. (2004). "Structure and evolution of transcriptional regulatory networks." Curr Opin Struct Biol **14**(3): 283-91.
- Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-12.
- Barabasi, A. L., H. Jeong, et al. (2002). "Evolution of the social network of scientific collaborations." Physica A **311**((3-4)): 590--614.
- Basso, K., A. Margolin, et al. (*Submitted*). "Reverse engineering of regulatory networks in human B cells." Nat Genet.
- Beirlant, J., E. Dudewicz, et al. (1997). "Nonparametric entropy estimation: An overview." Int. J. Math. Stat. Sci. **6**(1): 17-39.
- Bethe, H. (1935). "Statistical Theory of Superlattices." Proc. Roy. Soc. London A **150**: 552.
- Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." Pac Symp Biocomput: 418-29.
- Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. Learning from Data: Artificial Intelligence and Statistics. D. a. L. Fisher, H. New York, Springer-Verlag: 121-130.
- Cooper, G. F., and Herskovits, E. (1992). "A Bayesian method for the induction of probabilistic networks from data." Machine Learning **9**: 309-347.
- Cover, T. M. and J. A. Thomas (1991). Elements of Information Theory. New York, John Wiley & Sons.
- de la Fuente, A., P. Brazhnik, et al. (2002). "Linking the genes: inferring quantitative gene networks from microarray data." Trends Genet **18**(8): 395-8.
- Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.
- Erdos, P. and A. Renyi (1959). "On Random Graphs." Publ. Math. Debrecen **6**: 290-297.
- Fernandez, P. C., S. R. Frank, et al. (2003). "Genomic targets of the human c-Myc protein." Genes Dev **17**(9): 1115-29.
- Friedman, N. (2004). "Inferring cellular networks using probabilistic graphical models." Science **303**(5659): 799-805.
- Friedman, N. and G. Elidan (2004). LibB 2.1, <http://www.cs.huji.ac.il/labs/compbio/LibB/>.
- Gardner, T. S., D. di Bernardo, et al. (2003). "Inferring genetic networks and identifying compound mode of action via expression profiling." Science **301**(5629): 102-5.
- Gat-Viks, I. and R. Shamir (2003). "Chain functions and scoring functions in genetic networks." Bioinformatics **19 Suppl 1**: i108-17.
- Giot, L., J. S. Bader, et al. (2003). "A protein interaction map of Drosophila melanogaster." Science **302**(5651): 1727-36.
- Hartemink, A. J., D. K. Gifford, et al. (2001). "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks." Pac Symp Biocomput: 422-33.

- Heckerman, D. A. (1999). A tutorial on learning with Bayesian Networks. Learning in Graphical Models. M. Jordan. Cambridge, MA, MIT Press.
- Hill, A. V. (1910). "The possible effect of the aggregation of the molecules in hemoglobin." J. Physiol. **40**: iv-vii.
- Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.
- Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**(5518): 929-34.
- Iossifov, I., M. Krauthammer, et al. (2004). "Probabilistic inference of molecular networks from noisy data sources." Bioinformatics **20**(8): 1205-13.
- Janes, E. T. (1957). "Information theory and statistical mechanics." Phys. Rev. **106**: 620--630.
- Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.
- Jeong, H., B. Tombor, et al. (2000). "The large-scale organization of metabolic networks." Nature **407**(6804): 651-4.
- Joe, H. (1997). Multivariate models and dependence concepts. Boca Raton, FL, Chapman & Hall.
- Kabashima, Y. and D. Saad (2001). The TAP approach to intensive and extensive connectivity systems. Advanced Mean Field Methods: Theory and Practice. M. Oppen and D. Saad. Cambridge, MA, MIT Press.
- Kikuchi, R. (1951). "A Theory of Cooperative Phenomena." Phys. Rev. **81**: 988.
- Klein, U., Y. Tu, et al. (2001). "Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells." J Exp Med **194**(11): 1625-38.
- Kraskov, A., H. Stoeckbauer, et al. (2004). "Estimating mutual information." Phys. Rev. E **69**(6): 066138.
- Lamb, J., S. Ramaswamy, et al. (2003). "A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer." Cell **114**(3): 323-34.
- Ma, S.-K. (1985). Statistical mechanics. Singapore, World Scientific.
- Mangan, S. and U. Alon (2003). "Structure and function of the feed-forward loop network motif." Proc Natl Acad Sci U S A **100**(21): 11980-5.
- Mendes, P., W. Sha, et al. (2003). "Artificial gene networks for objective comparison of analysis algorithms." Bioinformatics **19 Suppl 2**: II122-II129.
- Mezard, M. and G. Parizi (2001). "The Bethe lattice spin glass revisited." Eur. Phys. J. B **20**: 217.
- Middendorf, M., A. Kundaje, et al. (2004). "Predicting genetic regulatory response using classification." Bioinformatics **20 Suppl 1**: I232-I240.
- Nemenman, I. (2004). Information theory, multivariate dependence, and genetic network inference. Tech. Rep., KITP, UCSB: arXiv: q-bio/0406015.
- Nemenman, I. and W. Bialek (2002). "Occam factors and model-independent Bayesian learning of continuous distributions." Phys. Rev. E **65**: 026137.
- Nemenman, I., W. Bialek, et al. (2004). "Entropy and information in neural spike trains: Progress on the sampling problem." Phys. Rev. E **69**: 056111.
- Nemenman, I., F. Shafee, et al. (2002). Entropy and Inference, Revisited. Adv. Neural Inf. Proc. Syst., MIT Press.

- Nemenman, I. and N. Tishby (*Submitted*). An axiomatic approach to the theory of information processing in networks.
- Opper, M. and O. Winther (2001). From naive mean field theory to the TAP equations. Advanced mean field methods: theory and practice. M. Opper and D. Saad. Cambridge, MA, MIT Press.
- Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA, Morgan Kaufmann Publishers, Inc.
- Rice, J. J. and G. Stolovitzky (2004). "Making the most of it: pathway reconstruction and integrative simulation using the data at hand." Biosilico(2): 70-77.
- Rice, J. J., Y. Tu, et al. (2004). "Reconstructing biological networks using conditional correlation analysis." Bioinformatics.
- Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.
- Steffen, M., A. Petti, et al. (2002). "Automated modelling of signal transduction networks." BMC Bioinformatics **3**(1): 34.
- Strong, S. P., R. Koberle, et al. (1998). "Entropy and information in neural spike trains." Phys. Rev. Lett. **80**(1): 197-200.
- Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.
- Tegner, J., M. K. Yeung, et al. (2003). "Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling." Proc Natl Acad Sci U S A **100**(10): 5944-9.
- Wiggins, C. and I. Nemenman (2003). "Process pathway inference via time series analysis." Experimental Mechanics **43**(3): 361--370.
- Yedidia, J. (2001). An idiosyncratic journey beyond mean field theory. Advanced Mean Field Methods: Theory and Practice. M. Opper and D. Saad. Cambridge, MA, MIT Press.
- Yu, J., A. V. Smith, et al. (2002). Using Bayesian Network Inference Algorithms to Recover Molecular Genetic Regulatory Networks. 3rd International Conference on Systems Biology.
- Zeller, K. I., A. G. Jegga, et al. (2003). "An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets." Genome Biol **4**(10): R69.

Figure 1

The mean absolute percent error in estimating mutual information for bivariate normal densities is compared to the percent of errors in ranking the relative mutual information values for randomly sampled pairs for which the distribution with the lower true MI value is within between 70% and 99% of the distribution with the higher value. MI estimation error (dashed blue line) is highly sensitive to the choice of Gaussian kernel width used by the estimator and grows rapidly for non-optimal parameter choices. However, due to similar bias for distributions with close MI values, the error in ranking pairs of distributions (solid brown line) is much less sensitive to the choice of this parameter. These averages were produced using samples from 1,000 bivariate normal densities with a random uniformly distributed correlation coefficient $\rho \in [0.1, 0.9]$, such that $\bar{I} = \frac{1}{2} \log(1 - \rho^2)$. This results in a distribution of MI values that closely resembles that of the real microarray data (not shown).

Figure 2

Examples of the Data Processing Inequality.

(a) g_1 , g_2 , and g_3 , are connected in a linear chain relationship. Although all six gene pairs will likely have enriched mutual information, the DPI will infer the most likely path of information flow. For example, $g_1 \leftrightarrow g_3$ will be eliminated because $I(g_1, g_2) > I(g_1, g_3)$ and $I(g_2, g_3) > I(g_1, g_3)$. $g_2 \leftrightarrow g_4$ will be eliminated because $I(g_2, g_3) > I(g_2, g_4)$ and $I(g_3, g_4) > I(g_2, g_4)$. $g_1 \leftrightarrow g_4$ will be eliminated in two ways: first, because $I(g_1, g_2) > I(g_1, g_4)$ and $I(g_2, g_4) > I(g_1, g_4)$, and then because $I(g_1, g_3) > I(g_1, g_4)$ and $I(g_3, g_4) > I(g_1, g_4)$. **(b)** If the underlying interactions form a tree (and MI can be measured without errors), ARACNE will reconstruct the network exactly by removing all false candidate interactions (dashed blue lines) and retaining all true interactions (solid black lines).

Figure 3

Reconstruction of the three-gene *S. cerevisiae* galactose regulatory network. **(a)** Eight distinct adirectional topologies must be evaluated by Bayesian Networks. One incorrect configuration (#1, black circles) has a higher score, and another (#6, gray circles) has the same score as the correct configuration (#4, light blue circles). **(b)** ARACNE evaluates mutual informations of three edges (shown near each edge) and correctly removes the edge between Gal4 and Gal80. Results were calculated using data from 52 Affymetrix GeneChips provided by (Hartemink, Gifford et al. 2001).

Figure 4

Topology of the 100 gene regulatory networks proposed by (Mendes, Sha et al. 2003). Blue/red edges correspond to activation/inhibition. For the Erdős-Rényi topology **(a)** each gene is equally likely to be connected to every other gene, while the scale-free topology **(b)** is characterized by large interaction hubs with many connections.

Figure 5

Precision vs. Recall for 1,000 samples generated from the Mendes networks. **(a)** Erdős-Rényi network topology. **(b)** Scale-free topology. PRCs for ARACNE are consistently better than for the other algorithms, and its precision reaches ~100% while maintaining high recall.

Figure 6

Distribution of mutual information for different lengths of the shortest path between genes for the **(a)** Erdős-Rényi topology and **(b)** scale-free topology. Here we plot the log of the empirical probability that MI for a given separation between genes is above some value (in nats) marked on the horizontal axis. For both topologies, high MI values are significantly more probable for closer genes. Statistical significance thresholds of 10^{-5} for the background MI distribution, corresponding to $I_0 = 0.0175$ nats, is marked on each graph. As shown, this threshold retains a large number of indirect candidate interactions, and there is no threshold that would be able to separate indirect and direct interactions; a threshold that eliminates most of the former (red arrows) also eliminates the majority of the latter. This severely degrades performance of RNs. **(b, inset)** Expanded log-log view of the MI distribution for 934 gene pairs with 3 or more intermediaries and the background distribution computed by Monte Carlo. The curves are virtually indistinguishable, indicating that the background distribution can be used to obtain reliable estimates of statistical significance thresholds for filtering genes with higher degrees of connectivity.

Figure 7

The total number of inferred errors (false positives plus false negatives) is stable with respect to choice of Gaussian kernel width for the estimator, validating the previous observation that errors in ranking MI for a pair of variables is more stable than the MI estimation error with respect to changes in this parameter (Figure 1). The choice of kernel width for each number of samples that minimizes the mean absolute MI estimation error for bivariate Gaussian densities (indicated with diamonds) yields optimal or near optimal reconstruction of this network for all samples sizes. Moreover, performance of the algorithm degrades gracefully as the number of samples decreases. Results are calculated for a statistical significance threshold of 10^{-5} and a synthetic microarray size of 1,000 for the scale-free network topology. Similar results apply for the Erdős-Rényi topology (see supplemental).

Figure 8

The number of inferred errors are plotted as a function of the DPI tolerance, τ . Raising τ to a value of 0.2 results in a modest increase in false positives, while larger values of τ produce a much sharper increase. Results are calculated for a statistical significance threshold of 10^{-5} and a synthetic microarray size of 1,000 for the scale-free topology. Similar results apply for the Erdős-Rényi topology (see supplemental).

Figure 9

Removed pending journal publication.

Table 1

Recovery for varying numbers of samples generated from the Mendes networks, which contain an average of ~ 194 true interactions after self-loops and bidirectional edges are eliminated. Results are calculated using a p-value of 10^{-5} for ARACNE and Relevance Networks, yielding < 0.5 expected false positives for 4,950 potential interactions, and using a Dirichlet prior with equivalent sample size of one for Bayesian Networks (Hartemink, Gifford et al. 2001). Results are averaged over three network configurations for each topology. For all sample sizes ARACNE efficiently eliminates almost all false candidate interactions inferred by RNs, as indicated by the DPI sensitivity (calculated as the percent of false positives eliminated by the DPI), with minimal reduction in true positives, as indicated by the DPI precision (calculated as the percent of false positives removed out of the total number of edges removed by the DPI). Moreover, as the sample size decreases, the number of true connections inferred by ARACNE decays gracefully while the number of false positives remains very low, whereas the performance of Bayesian Networks degrades rapidly for smaller sample sizes.

Table 2

Removed pending journal publication.

Figure 1

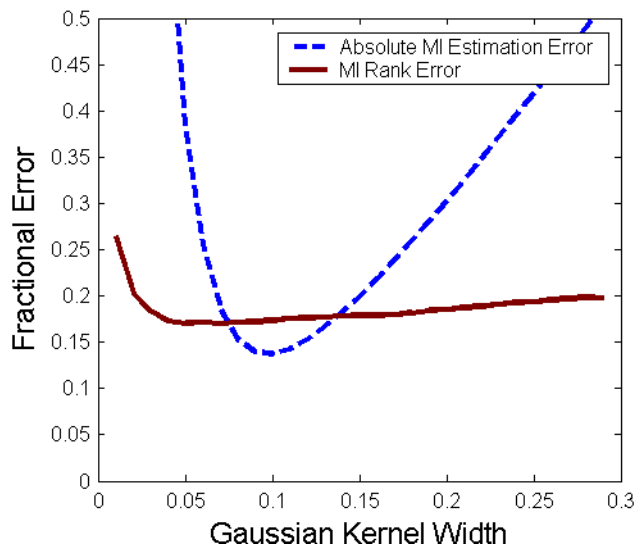


Figure 2

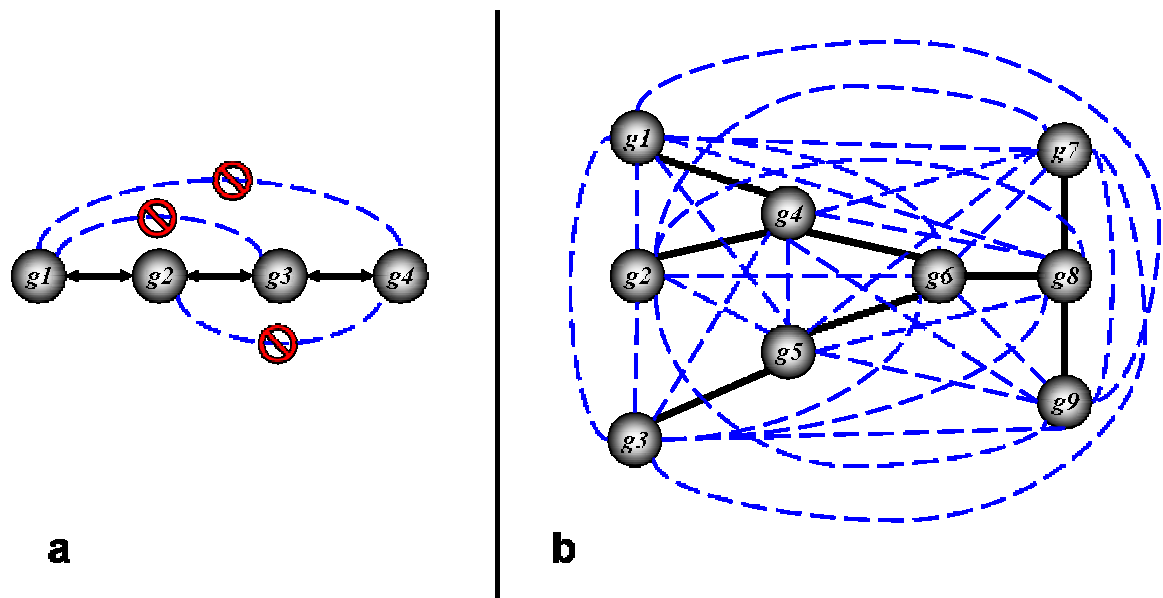


Figure 3

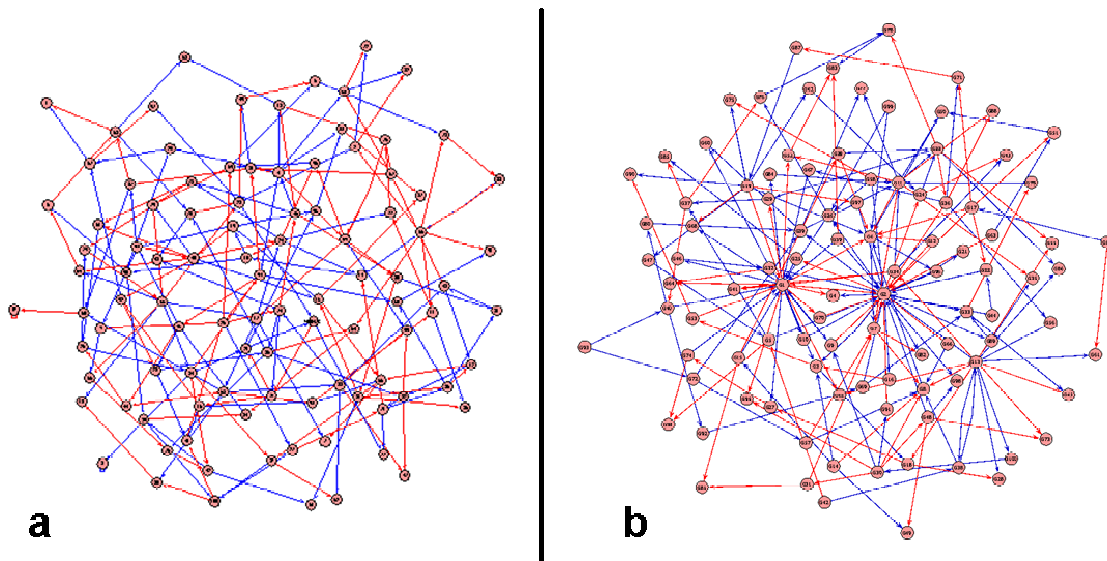


Figure 4

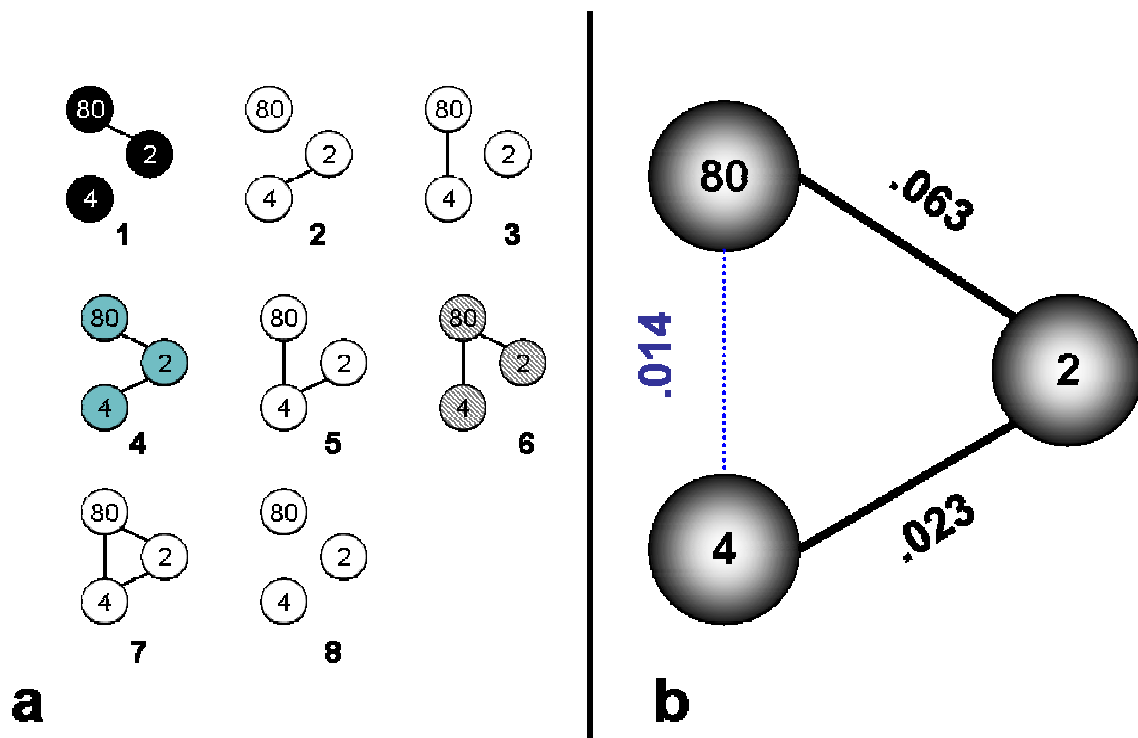


Figure 5

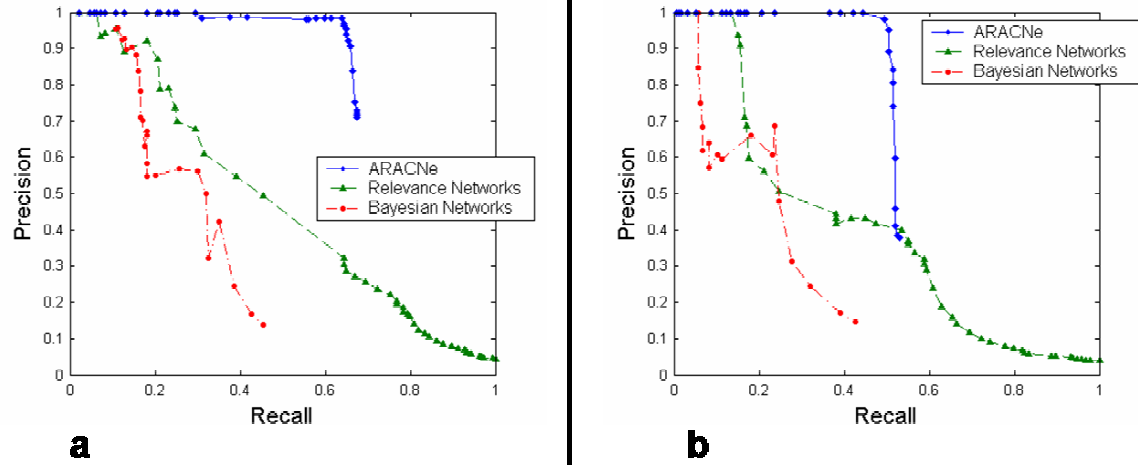


Figure 6

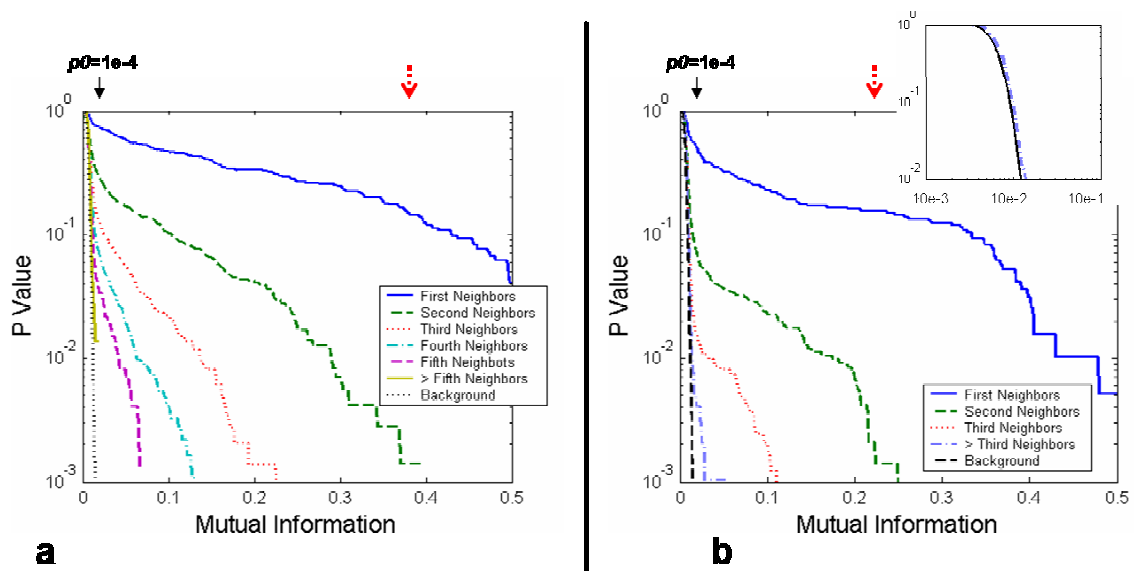


Figure 7

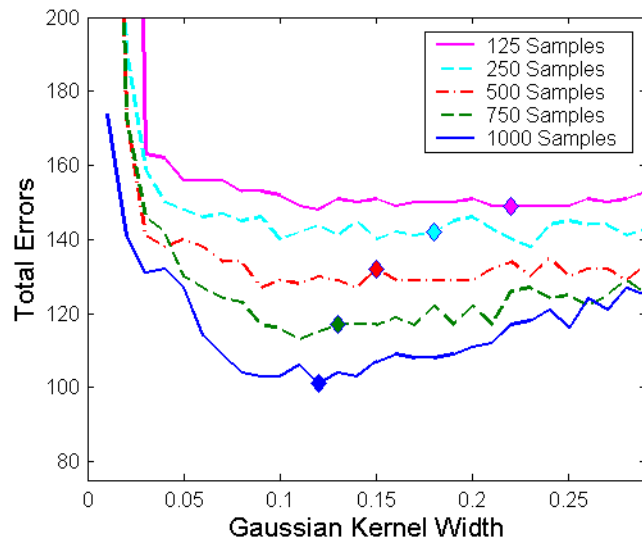


Figure 8

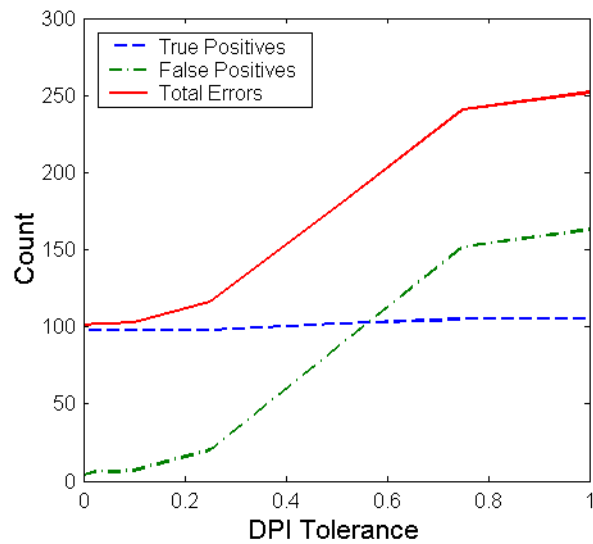


Figure 9

Removed pending journal publication.

Table 1***Erdős-Rényi Topology***

Num samples	ARACNE		Relevance Networks		DPI Sensitivity	DPI Precision	Bayesian Networks	
	N_{TP}	N_{FP}	N_{TP}	N_{FP}			N_{TP}	N_{FP}
1000	128.00	1.33	143.33	462.67	99.71%	96.78%	52.67	35.33
750	124.33	2.67	139.33	411.00	99.35%	96.46%	49.67	33.33
500	119.00	1.67	130.67	311.33	99.46%	96.37%	45.00	33.00
250	101.00	4.67	110.00	182.33	97.44%	95.18%	33.67	26.67
125	81.00	4.67	84.67	95.00	95.09%	96.10%	7.33	22.67

Scale-Free Topology

Num samples	ARACNE		Relevance Networks		DPI Sensitivity	DPI Precision	Bayesian Networks	
	N_{TP}	N_{FP}	N_{TP}	N_{FP}			N_{TP}	N_{FP}
1000	97.67	2.33	113.33	234.00	99.00%	93.67%	40.00	18.67
750	90.67	3.33	103.00	200.00	98.33%	94.10%	34.33	17.00
500	80.33	5.33	91.67	154.67	96.55%	92.95%	29.67	15.67
250	63.33	7.67	70.00	80.00	90.42%	91.56%	11.67	11.67
125	46.33	3.67	48.00	49.67	92.62%	96.50%	4.33	7.00

Table 2*Removed pending journal publication.*